ParaMetric: An Automatic Evaluation Metric for Paraphrasing

ParaMetric: An Automatic Evaluation Metric for Paraphrasing

Chris Callison-Burch
Center for Speech and Language Processing
Johns Hopkins University
3400 N. Charles St.
Baltimore, MD 21218

Trevor Cohn **Mirella Lapata**
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh, EH8 9LW

Abstract

We present ParaMetric, an automatic evaluation metric for data-driven approaches to paraphrasing. ParaMetric provides an objective measure of quality using a collection of multiple translations whose paraphrases have been manually annotated. ParaMetric calculates precision and recall scores by comparing the paraphrases discovered by automatic paraphrasing techniques against gold standard alignments of words and phrases within equivalent sentences. We report scores for several established paraphrasing techniques.

1 Introduction

Paraphrasing is useful in a variety of natural language processing applications including natural language generation, question answering, multi-document summarization and machine translation evaluation. These applications require paraphrases for a wide variety of domains and language usage. Therefore building hand-crafted lexical resources such as WordNet (Miller, 1990) would be far too laborious. As such, a number of data-driven approaches to paraphrasing have been developed (Lin and Pantel, 2001; Barzilay and McKeown, 2001; Barzilay and Lee, 2003; Pang et al., 2003; Quirk et al., 2004; Bannard and Callison-Burch, 2005). Despite this spate of research, no objective evaluation metric has been proposed.

In absence of a repeatable automatic evaluation, the quality of these paraphrasing techniques was gauged using subjective manual evaluations. Section 2 gives a survey of the various evaluation methodologies used in previous research. It has not been possible to directly compare paraphrasing techniques, because each one was evaluated using its own idiosyncratic experimental design. Moreover, because these evaluations were performed manually, they are difficult to replicate.

We introduce an automatic evaluation metric, called ParaMetric, which uses paraphrasing techniques to be compared and enables an evaluation to be easily repeated in subsequent research. ParaMetric utilizes data sets which have been annotated with paraphrases. ParaMetric compares automatic paraphrases against reference paraphrases.

In this paper we:

- Present a novel automatic evaluation metric for data-driven paraphrasing methods;
- Describe how manual alignments are created by annotating correspondences between words in multiple translations;
- Show how phrase extraction heuristics from statistical machine translation can be used to enumerate paraphrases from the alignments;
- Report ParaMetric scores for a number of existing paraphrasing methods.

2 Related Work

No consensus has been reached with respect to the proper methodology to use when evaluating paraphrase quality. This section reviews past methods for paraphrase evaluation.

Researchers usually present the quality of their automatic paraphrasing technique in terms of a subjective manual evaluation. These have used a variety of criteria. For example, Barzilay and McKeown (2001) evaluated their paraphrases by asking judges whether paraphrases were "approximately conceptually equivalent." Ibrahim et al. (2003) asked judges whether their paraphrases were "roughly interchangeable given the genre." Bannard and Callison-Burch (2005) replaced phrases with paraphrases in a number of

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 97–104
Manchester, August 2008

sentences and asked judges whether the substitutions "preserved meaning and remained grammatical." These subjective evaluations are rather vaguely defined and not easy to reproduce.

Others evaluate paraphrases in terms of whether they improve performance on particular tasks. Callison-Burch et al. (2006b) measure improvements in translation quality in terms of Bleu score (Papineni et al., 2002) and in terms of subjective human evaluation when paraphrases are integrated into a statistical machine translation system. Lin and Pantel (2001) manually judge whether a paraphrase might be used to answer questions from the TREC question-answering track. To date, no one has used task-based evaluation to compare different paraphrasing methods. Even if such an evaluation were performed, it is unclear whether the results would hold for a different task. Because of this, we strive for a general evaluation rather than a task-specific one.

Dolan et al. (2004) create a set of manual word alignments between pairs of English sentences. We create a similar type of data, as described in Section 4. Dolan et al. use heuristics to draw pairs of English sentences from a comparable corpus of newswire articles, and treat these as potential paraphrases. In some cases these sentence pairs are good examples of paraphrases, and in some cases they are not. Our data differs because it is drawn from multiple translations of the same foreign sentences. Barzilay (2003) suggested that multiple translations of the same foreign source text were a perfect source for "naturally occurring paraphrases" because they are samples of text which convey the same meaning but are produced by different writers. That being said, it may be possible to use Dolan et al's data toward a similar end. Cohn et al. (to appear) compares the use of the multiple translation corpus with the MSR corpus for this task.

The work described here is similar to work in summarization evaluation. For example, in the Pyramid Method (Nenkova et al., 2007) content units that are similar across human-generated summaries are hand-aligned. These can have alternative wordings, and are manually grouped. The idea of capturing these and building a resource for evaluating summaries is in the same spirit as our methodology.

3 Challenges for Evaluating Paraphrases Automatically

There are several problems inherent to automatically evaluating paraphrases. First and foremost, developing an exhaustive list of paraphrases for any given phrase is difficult. Lin and Pantel (2001) illustrate the difficulties that people have generating a complete list of paraphrases, reporting that they missed many examples generated by a system that were subsequently judged to be correct. If a list of reference paraphrases is incomplete, then using it to calculate precision will give inaccurate numbers. Precision will be falsely low if the system produces correct paraphrases which are not in the reference list. Additionally, recall is indeterminable because there is no way of knowing how many correct paraphrases exist.

There are further impediments to automatically evaluating paraphrases. Even if we were able to come up with a reasonably exhaustive list of paraphrases for a phrase, the acceptability of each paraphrase would vary depending on the context of the original phrase (Szpektor et al., 2007). While lexical and phrasal paraphrases can be evaluated by comparing them against a list of known paraphrases (perhaps customized for particular contexts), this cannot be naturally done for structural paraphrases which may transform whole sentences.

We attempt to resolve these problems by having annotators indicate correspondences in pairs of equivalent sentences. Rather than having people enumerate paraphrases, we asked that they perform the simper task of aligning paraphrases. After developing these manual "gold standard alignments" we can gauge how well different automatic paraphrases are at *aligning* paraphrases within equivalent sentences. By evaluating the performance of paraphrasing techniques at alignment, rather than at matching a list of reference paraphrases, we obviate the need to have a complete set of paraphrases.

We describe how sets of reference paraphrases can be extracted from the gold standard alignments. While these sets will obviously be fragmentary, we attempt to make them more complete by aligning *groups* of equivalent sentences rather than only pairs. The paraphrase sets that we extract are appropriate for the particular contexts. Moreover they may potentially be used to study structural paraphrases, although we do not examine that

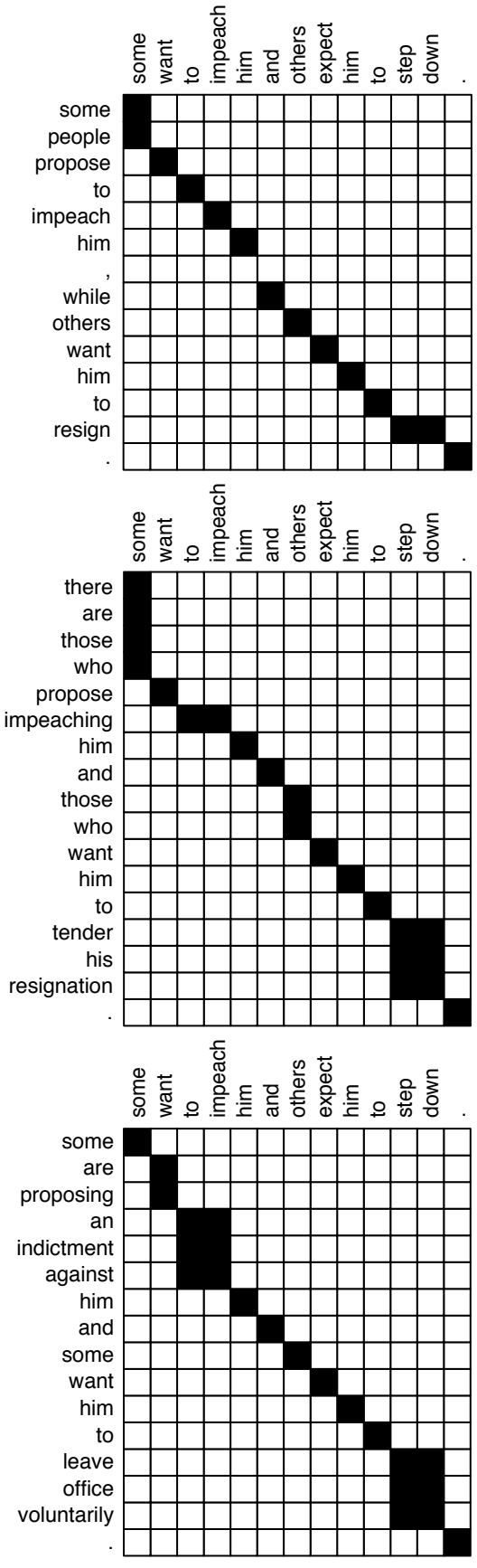

Figure 1: Pairs of English sentences were aligned by hand. Black squares indicate paraphrase correspondences.

in this paper.

4 Manually Aligning Paraphrases

We asked monolingual English speakers to align corresponding words and phrases across pairs of equivalent English sentences. The English sentences were equivalent because they were translations of the same foreign language text created by different human translators. Our annotators were instructed to align parts of the sentences which had the same meaning. Annotators were asked to prefer smaller one-to-one word alignments, but were allowed to create one-to-many and many-to-many alignments where appropriate. They were given a set of annotation guidelines covering special cases such as repetition, pronouns, genitives, phrasal verbs and omissions (Callison-Burch et al., 2006a). The manual correspondences are treated as *gold standard alignments*.

We use a corpus that contains eleven English translations of Chinese newswire documents, which were commissioned from different translation agencies by the Linguistics Data Consortium[1]. The data was created for the Bleu machine translation evaluation metric (Papineni et al., 2002), which uses multiple translations as a way of capturing allowable variation in translation. Whereas the Bleu metric requires no further information, our method requires a one-off annotation to explicitly show which parts of the multiple translations constitute paraphrases.

The rationale behind using a corpus with eleven translations was that a greater number of translations would likely result in a greater number of paraphrases for each phrase. Figure 1 shows the alignments that were created between one sentence and three of its ten corresponding translations. Table 1 gives a list of non-identical words and phrases that can be paired by way of the word alignments. These are the *basic paraphrases* contained within the three sentence pairs. Each phrase has up to three paraphrases. The maximum number of paraphrases for a given span in each sentence is bounded by the number of equivalent sentences that it is paired with.

In addition to these basic paraphrases, longer paraphrases can also be obtained using the heuristic presented in Och and Ney (2004) for extracting phrase pairs (PP) from word alignments A, between a foreign sentence f_1^J and an English sen-

[1]See LDC catalog number 2002T01.

some	some people, there are those who
want	propose, are proposing
to impeach	an indictment against, impeaching
and	while
others	some, those who
expect	want
step down	resign, leave office voluntarily, tender his resignation

Table 1: Non-identical words and phrases which are identified as being in correspondence by the alignments in Figure 1.

tence e_1^I:

$$PP(f_1^J, e_1^I, A) = \{(f_j^{j+m}, e_i^{i+n}) :$$

$$\forall (i', j') \in A : j \leq j' \leq j+m \leftrightarrow i \leq i' \leq i+n$$

$$\wedge \exists (i', j') \in A : j \leq j' \leq j+m \wedge \leftrightarrow i \leq i' \leq i+n\}$$

When we apply the phrase extraction heuristic to aligned English sentences, we add the constraint $f_j^{j+m} \neq e_i^{i+n}$ to exclude phrases that are identical. This heuristic would allow "*some people propose to impeach him*," "*some are proposing an indictment against him*," and "*there are those who propose impeaching him*" to be extracted as paraphrases of "*some want to impeach him*." The heuristic extracts a total of 142 non-identical phrase pairs from the three sentences given in Figure 1.

For the results reported in this paper, annotators aligned *50 groups* of 10 pairs of equivalent sentences, for a total of *500 sentence pairs*. These were assembled by pairing the first of the LDC translations with the other ten (i.e. 1-2, 1-3, 1-4, ..., 1-11). The choice of pairing one sentence with the others instead of doing all pairwise combinations was made simply because the latter would not seem to add much information. However, the choice of using the first translator as the key was arbitrary.

Annotators corrected a set of automatic word alignments that were created using Giza++ (Och and Ney, 2003), which was trained on a total of 109,230 sentence pairs created from all pairwise combinations of the eleven translations of 993 Chinese sentences.

The average amount of time spent on each of the sentence pairs was 77 seconds, with just over eleven hours spent to annotate all 500 sentence pairs. Although each sentence pair in our data set was annotated by a single annotator, Cohn et al. (to appear) analyzed the inter-annotator agreement for randomly selected phrase pairs from the same corpus, and found inter-annotator agreement of $\hat{C} = 0.85$ over the aligned words and $\hat{C} = 0.63$ over the alignments between basic phrase pairs, where $\hat{C}$ is measure of inter-rater agreement in the style of Kupper and Hafner (1989).

5 ParaMetric Scores

We can exploit the manually aligned data to compute scores in two different fashions. First, we can calculate how well an automatic paraphrasing technique is able *to align* the paraphrases in a sentence pair. Second, we can calculate the *lower-bound on precision* for a paraphrasing technique and its *relative recall* by enumerating the paraphrases from each of the sentence groups. The first of these score types does not require groups of sentences, only pairs.

We calculate alignment accuracy by comparing the manual alignments for the sentence pairs in the test corpus with the alignments that the automatic paraphrasing techniques produce for the same sentence pairs. We enumerate all non-identical phrase pairs within the manually word-aligned sentence pairs and within the automatically word aligned sentence pairs using PP. We calculate the precision and recall of the alignments by taking the intersection of the paraphrases extracted from the manual alignments M, and the paraphrases extracted from a system's alignments S:

$$Align_{Prec} = \frac{\sum_{<e_1,e_2>\in C} |PP(e_1, e_2, S) \cap PP(e_1, e_2, M)|}{\sum_{<e_1,e_2>\in C} |PP(e_1, e_2, S)|}$$

$$Align_{Recall} = \frac{\sum_{<e_1,e_2>\in C} |PP(e_1, e_2, S) \cap PP(e_1, e_2, M)|}{\sum_{<e_1,e_2>\in C} |PP(e_1, e_2, M)|}$$

Where e_1, e_2 are pairs of English sentence from the test corpus.

Measuring a paraphrasing method's performance on the task of *aligning* the paraphrases is somewhat different than what most paraphrasing methods do. Most methods produce a list of paraphrases for a given input phrase, drawing from a large set of rules or a corpus larger than our small test set. We therefore also attempt to measure precision and recall by comparing the set of

paraphrases that method M produces for phrase p that occurs in sentence s. We denote this set as $para_M(p, s)$, where s is an optional argument for methods that constrain their paraphrases based on context.

Our reference sets of paraphrases are generated in a per group fashion. We enumerate the reference paraphrases for phrase p in sentence s in group G as

$$para_{REF}(p_1, s_1, G) =$$

$$\{p_2 : \forall(p_1, p_2) \in \sum_{<s_1,s_2,A>\in G} PP(s_1, s_2, A)\}$$

The maximum size of this set is the number of sentence pairs in G. Because this set of reference paraphrases is incomplete, we can only calculate a lower bound on the precision of a paraphrasing method and its recall relative to the reference paraphrases. We call these $LB\text{-}Precision$ and $Rel\text{-}Recall$ and calculate them as follows:

$$LB\text{-}Precision =$$

$$\sum_{<s,G>\in C} \sum_{p\in s} \frac{|para_M(p, s) \cap para_{REF}(p_1, s, G)|}{|para_M(p, s)|}$$

$$Rel\text{-}Recall =$$

$$\sum_{<s,G>\in C} \sum_{p\in s} \frac{|para_M(p, s) \cap para_{REF}(p_1, s, G)|}{|para_{REF}(p_1, s, G)|}$$

For these metrics we require the test corpus C to be a held-out set and restrict the automatic paraphrasing techniques from drawing paraphrases from it. The idea is instead to see how well these techniques are able to draw paraphrases from the other sources of data which they would normally use.

6 Paraphrasing Techniques

There are a number of established methods for extracting paraphrases from data. We describe the following methods in this section and evaluate them in the next:

- Pang et al. (2003) used *syntactic alignment* to merge parse trees of multiple translations,
- Quirk et al. (2004) treated paraphrasing as *monolingual statistical machine translation*,
- Bannard and Callison-Burch (2005) used *bilingual parallel corpora* to extract paraphrases.

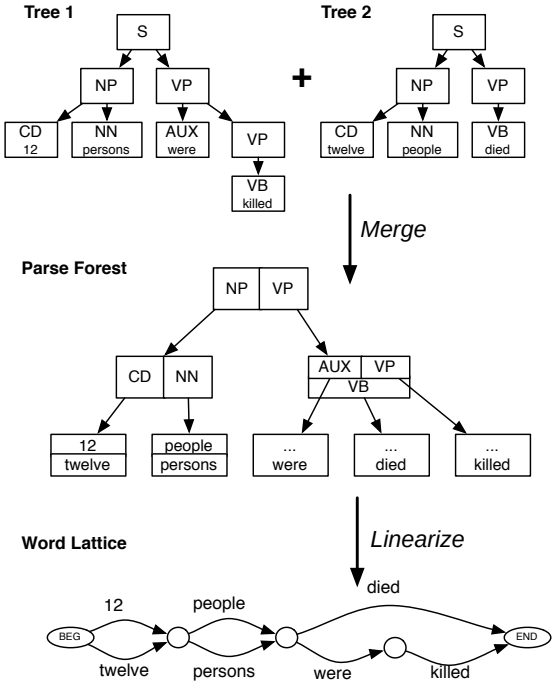

Figure 2: Pang et al. (2003) created word graphs by merging parse trees. Paths with the same start and end nodes are treated as paraphrases.

Pang et al. (2003) use multiple translations to learn paraphrases using a syntax-based alignment algorithm, illustrated in Figure 2. Parse trees were merged into forests by grouping constituents of the same type (for example, the two NPs and two VPs are grouped). Parse forests were mapped onto finite state word graphs by creating alternative paths for every group of merged nodes. Different paths within the resulting word lattice are treated as paraphrases of each other. For example, in the word lattice in Figure 2, *people were killed*, *persons died*, *persons were killed*, and *people died* are all possible paraphrases of each other.

Quirk et al. (2004) treated paraphrasing as "monolingual statistical machine translation." They created a "parallel corpus" containing pairs of English sentences by drawing sentences with a low edit distance from news articles that were written about the same topic on the same date, but published by different newspapers. They formulated the problem of paraphrasing in probabilistic terms in the same way it had been defined in the statistical machine translation literature:

$$\begin{aligned} \hat{e_2} &= \arg\max_{e_2} p(e_2|e_1) \\ &= \arg\max_{e_2} p(e_1|e_2)p(e_2) \end{aligned}$$

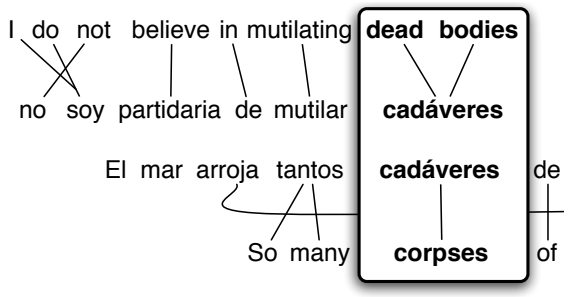

Figure 3: Bannard and Callison-Burch (2005) extracted paraphrases by equating English phrases that share a common translation.

	Parallel Corpora	Syntactic Alignment	Monolingual SMT
$Align_{Prec}$	.62	.65	.73
$Align_{Recall}$	.11	.10	.46
$LB\text{-}Precision$	.14	.33	.68
$Rel\text{-}Recall$	.07	.03	.01

Table 2: Summary results for scoring the different paraphrasing techniques using our proposed automatic evaluations.

Where $p(e_1|e_2)$ is estimated by training word alignment models over the "parallel corpus" as in the IBM Models (Brown et al., 1993), and phrase translations are extracted from word alignments as in the Alignment Template Model (Och, 2002).

Bannard and Callison-Burch (2005) also used techniques from statistical machine translation to identify paraphrases. Rather than drawing pairs of English sentences from a comparable corpus, Bannard and Callison-Burch (2005) used bilingual parallel corpora. They identified English paraphrases by pivoting through phrases in another language. They located foreign language translations of an English phrase, and treated the other English translations of those foreign phrases as potential paraphrases. Figure 3 illustrates how a Spanish phrase can be used as a point of identification for English paraphrases in this way. Bannard and Callison-Burch (2005) defined a paraphrase probability $p(e_2|e_1)$ in terms of the translation model probabilities $p(f|e_1)$ and $p(e_2|f)$. Since e_1 can translate as multiple foreign language phrases, they sum over f, and since multiple parallel corpora can be used they summed over each parallel corpus C:

$$\begin{aligned} \hat{e_2} &= \arg\max_{e_2 \neq e_1} p(e_2|e_1) \\ &\approx \arg\max_{e_2 \neq e_1} \sum_{C} \sum_{f \ in \ C} p(f|e_1) p(e_2|f) \end{aligned}$$

7 Comparing Paraphrasing Techniques with ParaMetric

7.1 Training data for word alignments

In order to calculate $Align_{Prec}$ and $Align_{Recall}$ for the different paraphrasing techniques, we had them automatically align the 500 manually aligned sentence pairs in our test sets.

Bo Pang provided syntactic alignments for the 500 sentence pairs. The word lattices combine the groups of sentences. When measuring alignment quality, we took pains to try to limit the extracted phrase pairs to those which occurred in each sentence pair, but we acknowledge that our methodology may be flawed.

We created training data for the monolingual statistical machine translation method using all pairwise combination of eleven English translations in LDC2002T01. All combinations of the eleven translations of the 993 sentences in that corpus resulted in 109,230 sentence pairs with 3,266,769 words on each side. We used this data to train an alignment model, and applied it to the 500 sentence pairs in our test set.

We used the parallel corpus method to align each pair of English sentences by creating intermediate alignments through their Chinese source sentences. The bilingual word alignment model was trained on a Chinese-English parallel corpus from the NIST MT Evaluation consisting of 40 million words. This was used to align the 550 Chinese-English sentence pairs constructed from the test set.

7.2 Training data for precision and recall

Each of the paraphrasing methods generated paraphrases for $LB\text{-}Precision$ and $Rel\text{-}Recall$ using larger training sets of data than for the alignments. For the syntax-based alignment method, we excluded the 50 word lattices corresponding to the test set. We used the remaining 849 lattices for the LDC multiple translation corpus. For the monolingual statistical machine translation method, we downloaded the Microsoft Research Paraphrase Phrase Table, which contained paraphrases for nearly 9 million phrases, gener-

ated from the method described in Quirk et al. (2004). For the parallel corpus method, we derived paraphrases from the entire Europarl corpus, which contains parallel corpora between English and 10 other languages, with approximately 30 million words per language. We limited both the Quirk et al. (2004) and the Bannard and Callison-Burch (2005) paraphrases to those with a probability greater than or equal to 1%.

7.3 Results

Table 2 gives a summary of how each of the paraphrasing techniques scored using the four different automatic metrics. The precision of their alignments was in the same ballpark, with each paraphrasing method reaching above 60%. The monolingual SMT method vastly outstripped the others in terms of recall and therefore seems to be the best on the simplified task of aligning paraphrases within pairs of equivalent sentences.

For the task of generating paraphrases from unrestricted resources, the monolingual SMT method again had the highest precision, although time time its recall was quite low. The 500 manually aligned sentence pairs contained 14,078 unique paraphrases for phrases of 5 words or less. The monolingual SMT method only posited 230 paraphrases with 156 of them being correct. By contrast, the syntactic alignment method posited 1,213 with 399 correct, and the parallel corpus method posited 6,914 with 998 correct. Since the reference lists are incomplete by their very nature, the $LB\text{-}Precision$ score gives a lower-bound on the precision, and the $Rel\text{-}Recall$ gives recall only with respect to the partial list of paraphrases.

Table 3 gives the performance of the different paraphrasing techniques for different phrase lengths.

8 Conclusions

In this paper we defined a number of automatic scores for data-driven approaches to paraphrasing, which we collectively dub "ParaMetric". We discussed the inherent difficulties in automatically assessing paraphrase quality. These are due primarily to the fact that it is exceedingly difficult to create an exhaustive list of paraphrases. To address this problem, we introduce an artificial task of aligning paraphrases within pairs of equivalent English sentences, which guarantees accurate precision and recall numbers. In order to measure alignment quality, we create a set of gold standard alignments. While the creation of this data does require some effort, it seems to be a manageable amount, and the inter-annotator agreement seems reasonable.

Since alignment is not perfectly matched with what we would like automatic paraphrasing techniques to do, we also use the gold standard alignment data to measure a lower bound on the precision of a method's paraphrases, as well as its recall relative to the limited set of paraphrases. Future studies should examine how well these scores rank different paraphrasing methods when compared to human judgments. Follow up work should investigate the number of equivalent English sentences that are required for reasonably complete lists of paraphrases. In this work we aligned sets of eleven different English sentences, but we acknowledge that such a data set is rare and might make it difficult to port this method to other domains or languages.

The goal of this work is to develop a set of scores that both allows different paraphrasing techniques to be compared objectively and provides an easily repeatable method for automatically evaluating paraphrases. This has hitherto not been possible. The availability of an objective, automatic evaluation metric for paraphrasing has the potential to impact research in the area in a number of ways. It not only allows for the comparison of different approaches to paraphrasing, as shown in this paper, but also provides a way to tune the parameters of a single system in order to optimize its quality.

Acknowledgments

The authors are grateful to Bo Pang for providing the word lattices from her method, to Stefan Riezler for his comments on an early draft of this paper, and to Michelle Bland for proofreading. This work was supported by the National Science Foundation under Grant No. 0713448. The views and findings are the authors' alone.

References

Bannard, Colin and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan.

Barzilay, Regina and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach us-

	$Align_{Prec}$			$Align_{Recall}$			$LB\text{-}Precision$			$Rel\text{-}Recall$		
	Parallel Corpora	Syntactic Alignment	Monolingual SMT	Parallel Corpora	Syntactic Alignment	Monolingual SMT	Parallel Corpora	Syntactic Alignment	Monolingual SMT	Parallel Corpora	Syntactic Alignment	Monolingual SMT
Length = 1	.54	.48	.64	.24	.18	.56	.15	.25	.59	.20	.16	.02
Length ≤ 2	.56	.56	.69	.19	.13	.52	.15	.31	.66	.18	.10	.03
Length ≤ 3	.59	.60	.71	.14	.12	.49	.15	.32	.66	.13	.06	.02
Length ≤ 4	.60	.63	.72	.12	.11	.48	.14	.33	.68	.09	.04	.01
Length ≤ 5	.62	.65	.73	.11	.10	.46	.14	.33	.68	.07	.03	.01

Table 3: Results for paraphrases of continuous subphrases of various lengths.

ing multiple-sequence alignment. In *Proceedings of HLT/NAACL-2003*, Edmonton, Alberta.

Barzilay, Regina and Kathleen McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL-2001)*.

Barzilay, Regina. 2003. *Information Fusion for Mutlidocument Summarization: Paraphrasing and Generation*. Ph.D. thesis, Columbia University, New York.

Brown, Peter, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Callison-Burch, Chris, Trevor Cohn, and Mirella Lapata. 2006a. Annotation guidelines for paraphrase alignment. Tech report, University of Edinburgh.

Callison-Burch, Chris, Philipp Koehn, and Miles Osborne. 2006b. Improved statistical machine translation using paraphrases. In *Proceedings of HLT/NAACL-2006*, New York, New York.

Cohn, Trevor, Chris Callison-Burch, and Mirella Lapata. to appear. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*.

Dolan, Bill, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th International Conference on Computational Linguistics*.

Ibrahim, Ali, Boris Katz, and Jimmy Lin. 2003. Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the Second International Workshop on Paraphrasing (ACL 2003)*.

Kupper, Lawrence L. and Kerry B. Hafner. 1989. On assessing interrater agreement for multiple attribute responses. *Biometrics*, 45(3):957–967.

Lin, Dekang and Patrick Pantel. 2001. Discovery of inference rules from text. *Natural Language Engineering*, 7(3):343–360.

Miller, George A. 1990. Wordnet: An on-line lexical database. *Special Issue of the International Journal of Lexicography*, 3(4).

Nenkova, Ani, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, 4(2).

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Och, Franz Josef and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Och, Franz Josef. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen Department of Computer Science, Aachen, Germany.

Pang, Bo, Kevin Knight, and Daniel Marcu. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of HLT/NAACL-2003*, Edmonton, Alberta.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, Pennsylvania.

Quirk, Chris, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, Barcelona, Spain.

Szpektor, Idan, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-2007)*, Prague, Czech Republic.